Castles in the Clouds: The Irrelevance of Vertical Scales for Most Practical Concerns

Castles in the clouds: The irrelevance of vertical scales for most practical concerns

Andrew Ho, Harvard Graduate School of Education

I imagine our community of modern testing experts as Ph.D.-level designers of the snap-together children's blocks known as Legos. We lovingly construct versatile pieces meant to cohere elegantly into castles and starships, with multistep instruction manuals detailing their appropriate use. And we sit back and shake our heads as Lego users neglect our instructions and build, instead, nondescript multicolor towers and, maybe, an occasional car. We think we are selling castles and starships, but our product sells less for its intended purposes than its flexibility; its perceived utility exists in the eyes of beholders.

Briggs and Peck (this issue, hereafter, B&P) propose what amounts to an elaborate Lego construction kit—this is my analogy for their recipe for defensible vertical scales. These are instructions for castles in the clouds. Their fundamental error is neither of construction nor aspiration. These are nice castles, and they are mostly well designed. Their error is one of misperception. They have made an insufficient case that anyone wants or needs castles over and above, say, cars. In this four-part response, I first argue that, contrary to the way that B&P motivate their castles, the difference between castles and cars is not a tension that needs to be resolved for any practical reason. I then argue that their proposed solution does not resolve this tension anyways—castles do not solve problems that cars do. To stretch the metaphor, I next argue that their focus on castles may risk the design of cars. I close by encouraging a focus back on frameworks for improving the use of test scores for practical purposes. Let's facilitate better use of cars as well as enabling castles.

An inconsequential disconnect: How "growth" is like "irony"

Briggs and Peck motivate their learning-progression approach to growth by noting that it resolves a tension between two different conceptions of growth. Extending Lord (1967) and Holland

and Rubin (1983), B&P imagine the evaluation of two teachers based upon the pretest-posttest scores of the teachers' students. B&P distinguish between comparing teachers using absolute gain scores (how different are students from their pretest scores, on average?) and residual gain scores[1] (how different are students from their predicted posttest scores, on average?). They note, without evidence, that many educational stakeholders assume that teacher effectiveness measures are based on gain scores, not residual gain scores. They also criticize the misleading use of the term, "growth," in research reports and testing manuals. I agree that these misperceptions and inaccuracies exist. What the B&P argument is missing is any compelling evidence to suggest that these misconceptions or inaccuracies are actually consequential.

The B&P lament about stakeholder misunderstanding of "growth," feels to me a lot like an English teacher's lament about the English-speaking public's misunderstanding of the term "irony." The first Merriam-Webster dictionary definition of irony (2015) is similar to that of sarcasm, "the use of words to express something other than and especially the opposite of the literal meaning." The more common use of the word irony is also sanctioned by Merriam-Webster (2015) and refers broadly, even vaguely, to "incongruity between the actual result of a sequence of events and the normal or expected result." While I agree that the imprecise use of the word "irony" by, say, the pop singer Alanis Morissette, in her song titled, "Ironic," is lamentable, I'm hard-pressed to cite examples of practical consequences of inaccurate use of the term. Misuses of this sort are less inaccurate than ambiguous and imprecise. Similarly, B&P make an incomplete case that stakeholder failure to communicate or comprehend a precise definition of growth is ultimately consequential.

In Table 1, I provide a framework within which B&P could have clarified their argument. What problem does incoherence between growth definition and growth interpretation actually pose, and how

[1] Following Castellano & Ho (2013, 2015), I prefer to call these "conditional status" scores at the student level and "aggregate conditional status" scores at the teacher level. I use the more colloquial "residual gain" term to highlight the confusion of residual gains with gains that B&P are attempting to resolve.

should we resolve it? Let's say that we design a teacher evaluation system based on residual gain scores. If a parent reads a report that says a teacher's "growth" score is 95, and the parent assumes it is based on a gain score, so what? If the parent is interested in the best estimate of teacher "value added" given only longitudinal student score histories, then residual-gain-based approaches are a more direct way to answer this question (whether this is a sufficiently good approach for the metric to have consequences is another matter). As long as the metric answers the parent's intended question, the only problem is that the parent misunderstands the basis for the metric. Trying to "fix" this is like interrupting a conversation between two people who are both using—and mutually understanding—the imprecise definition of the word, "ironic," just to note that there is another definition. So what?

The structure of the B&P argument only supports their ultimate proposal if they were to attack residual gains as an incorrect way to evaluate teachers. Without this attack, the way to resolve incoherence is not to design defensible vertical scales (the B&P proposal) but to teach parents that residual gains, not gain scores, are the basis for evaluating teachers that best answers their questions. In Table 1, I identify this as a hypothetical evaluator's solution to the incoherence. Throughout their paper, B&P take for granted that stakeholder intuition about gains suffices to warrant widespread adoption of gain-based growth measures. They err through omission, by neglecting that residual-gain-based metrics are appropriate for answering many practical questions. They could have argued more directly that stakeholder intuition about growth represents an opportunity to design metrics that meet that intuition, for formative purposes. I agree with this sentiment, but it neither resolves nor is relevant to the misrepresentation of appropriately designed residual-gain metrics as gain-based, a misrepresentation that I believe is largely inconsequential.

Legos can make cars as well as castles

Ironically, B&P raise Lord's paradox but do not take its implications as far as they should. The implications of the Holland and Rubin (1983) resolution of the paradox for gain scores is that gain scores

answer a different question (in which classroom is there a greater average gain?) than residual gains do (in which classroom is there a greater average score than expected given pretest scores?). The B&P argument seems to frame the two questions as a difference to be resolved, when instead it is just a difference, period. If B&P succeed in creating a defensible vertical scale, that's great, but the two questions remain different.

I like to resolve Lord's paradox with the observation that the residual-gain-based approach is like fitting a regression line for each group, whereas the gain-based approach is equivalent to constraining the slope parameters of the regression lines to 1. Like any other parameter constraint, this can be justified judgmentally and theoretically, whereas freeing the parameter for estimation is more exploratory and empirical. This perspective on Lord's paradox allows us to observe that the residual gain and gain-based approaches will only be equivalent of the slope of the empirical regression line for each group is 1. This will only be the case if $s_Y = s_X/r_{XY}$, when the standard deviation of the posttest scores (s_Y) is greater than that of the pretest scores (s_X) by a factor of the inverse correlation ($1/r_{XY}$). This is not guaranteed by any vertical scaling method, it would just have to be a matter of fact, empirically. There is nothing particular about the B&P vertical scaling proposal that would increase the likelihood that the two answers to these questions would be similar.

Nor is it desirable that these two answers be similar. From my "parameter constraint" perspective on Lord's paradox, I can read the B&P proposal as implying that we should do the theoretical and judgmental work necessary to defend our decision to constrain the slope parameters of our regression lines to 1. It's hard to see why constraining slopes would be necessary for the purposes of evaluation, when we could just as easily estimate slopes empirically, without all of the work that B&P propose. The B&P proposal for defensible vertical scales is a fine pursuit in and of itself, but following their proposal does not transform the two questions underlying Lord's Paradox into the same question, nor would it lead us to conclude that questions answered by gain scores hold unconditional relevance

over questions answered by residual gains. We can fit both regression lines (residual gains) and constrained regression lines (gain scores) to data, and they answer different questions. We can use Legos to build castles as well as cars.

Allow castles, but don't make it hard to build cars

I have argued that B&P err by motivating their vertical scales with accountability metrics, when, as one of the authors himself has noted elsewhere, most accountability metrics do not require vertical scales (Briggs & Domingue, 2013), and they are fairly robust even when equal-interval arguments are weak (Briggs & Betebenner, 2009; Castellano & Ho, 2015). When B&P observe that many vertical scales are underused, they conclude that this is because they are not sufficiently well constructed. I find a second conclusion more plausible: very few stakeholders find across-grade vertical scales useful for their purposes. I have argued that the B&P proposal neither resolves some consequential tension nor satiates some colossal, unmet demand.

Without this motivation, I believe the potential costs of using learning progressions for accountability metrics far outweigh the benefits. The B&P proposal, illustrated well in their Figure 3, is the definition of a narrowed curriculum. The authors concede this in their conclusion but believe it is nonetheless necessary for defensible growth measurement. As I have argued, the high standards to which B&P hold growth measurement are not necessary for evaluative purposes, and it is therefore not worth the cost of a narrowed curriculum.

It is clear to me that the usefulness of the B&P proposal in the near term is considerable but restricted to formative assessment, not evaluative assessment. I support the development of learning progressions in laboratories or otherwise controlled settings, and I support the use of these progressions by teachers who can understand the meaning of specific points along the progression and can thereby facilitate learning along it. However, for more general and flexible purposes, the domain definition of growth is a fine compromise that is less likely to narrow the curriculum (or as extremely, at least) while

also supporting grade-to-grade vertical scales that rarely see use in any case. Returning to the Lego metaphor, I see little reason to turn all Legos into custom castle construction kits if this would risk the construction of basic cars and towers

Increasing the demand for castles and raising the standards on cars

It is uncomfortable for me to argue for pragmatism when I hold great respect for the ideals that B&P espouse. As one who appreciates the difference between definitions of irony and enjoys building castles with Legos, I wish that we had more similarly minded company. However, I am less interested in advancing measurement for its own sake than in facilitating the use of measures for appropriate purposes, with the likelihood of benefiting their intended audiences. I see far greater promise for the B&P proposal for small-scale curricular use than I do for any large-scale policy use. And their measure of success should not be the achievement of equal-interval properties but the demonstration that using formative assessments developed in this fashion leads to better teaching and more learning than formative assessments developed conventionally.

I have yet to hear a convincing reason why equal-interval scales, over and above precise estimates on a unidimensional scale, are necessary to understand growth for practical formative purposes. For formative purposes, teachers and students need to understand what X means, what Y means, and how to best get from X to Y. I see little reason for them to care specifically about gains ($Y - X$) and even less reason to care specifically about comparing gains across students or average gains across teachers. If users do care about these comparisons, their interest is likely evaluative, not formative, and then residual/regression-based methods will provide more direct answers to their questions. Acceleration and deceleration along a vertical scale do require equal intervals and seem like they are important inferences to the extent that they are early warning indicators for future benchmarks, but, again, for these purposes, regression-based procedures are generally superior (Ho, 2014). I support the B&P proposal to pursue equal-interval properties from a scientific perspective,

however, again, I believe they have made an insufficient argument for their necessity for practical purposes.

The black box for measurement theory continues to be where the rubber meets the road: how users interpret scores and act upon them. While I am glad to have B&P and others advocating for better Legos and better instructions for their use, I also hope we can focus on two other lines of work: improving the construction of cars, and understanding how a car builder can be convinced to build castles. Rather than advocating for a product that holds no particular advantage for its most likely use, we should work to understand how users change, and can be caused to change, their primary use of tests from ranking, sorting, and selecting, to improving teaching and learning.

References

Briggs, D. C., & Betebenner, D. W. (2009). *Is growth in student achievement scale dependent?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*, 551-576.

Briggs, D. C., & Peck, P. A. (2015, this issue). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives, 13,* 75-99.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics, 40*, 35-68.

Ho, A. D. (2014). Accuracy, transparency, and incentives: Contrasting criteria for evaluating growth models. In R. W. Lissitz and H. Jiao (Eds.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness* (61-85). Charlotte, NC: Information Age Publishing.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Irony. (2015). In *Merriam-Webster.com*. Retrieved from http://www.merriam-webster.com/dictionary/irony

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304-305.

Table 1. The unclear consequences of incoherent definitions of growth.

		How do users think experts are defining growth?	
		Gain (probably most stakeholders?)	Residual Gain (probably most evaluators?)
How do experts define growth for their intended purposes?	Gain (formative)	Coherence (Briggs and Peck proposal?)	So what?
	Residual Gain (evaluative)	So what?	Coherence (An evaluator's proposal?)